Reliability Diagrams for Multicategory Probabilistic Forecasts

THOMAS M. HAMILL

Department of Soil, Crop, and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 10 December 1996, in final form 10 June 1997)

ABSTRACT

The most common method of verifying multicategory probabilistic forecasts such as are used in probabilistic quantitative precipitation forecasting is through the use of the ranked probability score. This single number description of forecast accuracy can never capture the multidimensional nature of forecast quality and does not inform the forecaster about the sources of forecast deficiencies. A new type of reliability diagram is developed here and applied to probabilistic quantitative precipitation forecasts from a university contest. This diagram is shown to potentially be useful in helping the forecaster to correct some errors in assigning the categorical probabilities.

1. Introduction

A particularly informative way of conveying quantitative precipitation forecast information is to express them as a vector of probabilities for a set of categories. For example, the Model Output Statistics, or MOS (Dallavalle et al. 1992) computes precipitation forecast probabilities for the categories $0 \leq V < 0.01$ in., $0.01 \leq V < 0.10$ in., $0.10 \leq V < 0.25$ in., $0.25 \leq V < 0.50$ in., $0.50 \leq V < 1.00$ in., and $1.00 \leq V$ in., where V is the verification (observed) amount (1.0 in. = 25.4 mm). In the Cornell University forecast contest (Hamill and Wilks 1995), we adopted this set of categories and ranked the competing forecasts using the ranked probability skill score, or RPSS (Wilks 1995), which is based on the ranked probability score (Epstein 1969; Murphy 1971; Daan 1985). This RPSS is a single number indicating the fractional improvement over a reference forecast. An RPSS of 0.0 indicates no difference in skill over the reference forecast, and an RPSS of 1.0 indicates a perfect forecast. In this contest, the reference forecast was persistence. As the RPSS is a single number, it distills the forecast performance to an understandable measure that is necessary for ranking competing forecasts. However, a forecaster seeking to understand how her forecasts are in error is not illuminated by the RPSS. Are the forecasts too sharp (specific), or biased? Are 25% of the forecasts on average below the 25th percentile of forecast distribution? The RPSS gives no such information. Hence, a new verification methodology was sought.

Corresponding author address: Dr. Thomas M. Hamill, NCAR/RAP, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: hamill@ucar.edu

The reliability diagram (Wilks 1995) is frequently used for assessing probability forecasts for binary events such as the probability of measurable precipitation. In the reliability diagram (Fig. 1), at a regular set of forecast probabilities such as 0%–100% percent in 10% intervals, the observed relative frequency of event occurrence is calculated and plotted. The diagram is well suited to assessing the ability to calibrate forecasts for binary events, but the use of this diagram with multicategory forecasts is not straightforward. Perhaps reliability diagrams could be generated by collapsing the multiple categories to binary probability forecasts. For example, the reliability for the event $V \geq 0.10$ in. (2.5 mm) can be calculated by subtracting from 1.0 the probabilities assigned to the categories $0 \leq V < 0.01$ in. and $0.01 \leq V < 0.10$ in. and tallying the observed relative frequency at the regular forecast probabilities. However, not all forecast verification questions can be answered after this simplification. For example, consider the user who is concerned only with the ability to assess categorical probabilities correctly for forecasts expected to verify in the 0.10–0.50 in. (2.5–12.7 mm) range. In such a circumstance, perhaps, another reliability diagram could be generated at the 0.50-in. threshold. However, the method for making inferences about the probability distribution from the two diagrams together is not clear. Further, a very large sample size is typically needed to sufficiently populate all of the probability bins; otherwise, the diagrams are noisy and useless.

The remainder of this article will demonstrate a technique for generating a new type of reliability diagram that does not require the conversion to a binary event; it is somewhat similar to the ''P–P plot'' (Wilks 1995). In doing so, the ''multicategory reliability diagram'' (hereafter called MCRD) is shown to be superior to simple reliability diagrams in assessing some probabil-

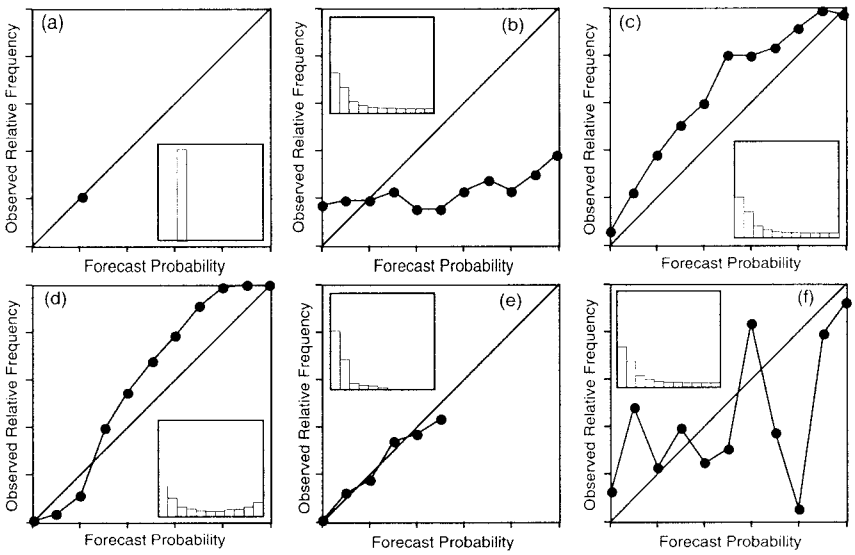

FIG. 1. Hypothetical reliability diagrams showing observed relative frequency as a function of forecast probabilities for (a) climatological forecasts, (b) forecasts exhibiting minimal resolution, (c) forecasts showing an underforecasting bias, (d) forecasts showing good resolution at the expense of reliability, (e) reliable forecasts of a rare event, and (f) verification dataset limited by small sample size. Inset boxes indicate frequency of use of the forecasts. Reprinted with permission from Wilks (1995).

ity forecast deficiencies in a simple manner. The MCRD will be shown to be intermediate in its complexity: as a multidimensional assessment of forecast quality, it cannot replace the RPSS for ranking competing forecasts. On the other hand, the MCRD collapses the dimensionality of the joint distribution of forecasts and observations, and thus cannot give a complete description of forecast quality. Section 2 will outline the mechanics of generating the MCRD; section 3 demonstrates an application of the diagram to forecasts from the Cornell University contest. Section 4 concludes.

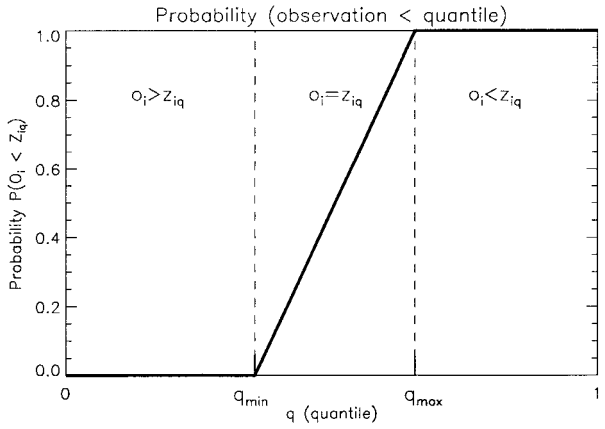

FIG. 2. Illustration of the allocation of probability for various quantiles given q_{min} and q_{max} are the lowest and highest quantiles where the observation and forecast have the same category.

2. Generating the multicategory reliability diagram

A conventional reliability diagram synthesizes whether probability forecasts for a binary (yes/no) event are well calibrated; at any given forecast probability of event occurrence, the observed relative frequency should be similar. The MCRD adopts the alternative goal of determining the average percentage of observations below prespecified quantiles (percentiles) of the forecast distribution. Hereafter, this will also be denoted as the ''calibration'' or ''reliability'' for a given quantile, though this terminology is perhaps nonstandard. To achieve perfect calibration, averaged over many forecast distributions, 25% of the observations should be below the 25th percentile of the distributions, 75% below the 75th percentile, and so on.

Suppose probability forecasts are to be made for J mutually exclusive and collectively exhaustive categories, such as the six MOS categories listed in section 1. On a given ith forecast day, a forecaster issues a probability forecast vector with elements y_{ij}, $j = 1, \ldots, J$, and $i = 1, \ldots, N$, where N is the total number of forecasts made (say, over many months of forecasting each day). For simplification, let us assume that probabilities for each category are rounded to the nearest 10%, for example, $\mathbf{y}_i = [0.7, 0.2, 0.1, 0, 0, 0]$.

This forecast vector is reexpressed as a vector of category numbers at a discrete set of quantiles, or percentiles within the distribution. The calibration will later be computed at these quantiles. For simplification, rather

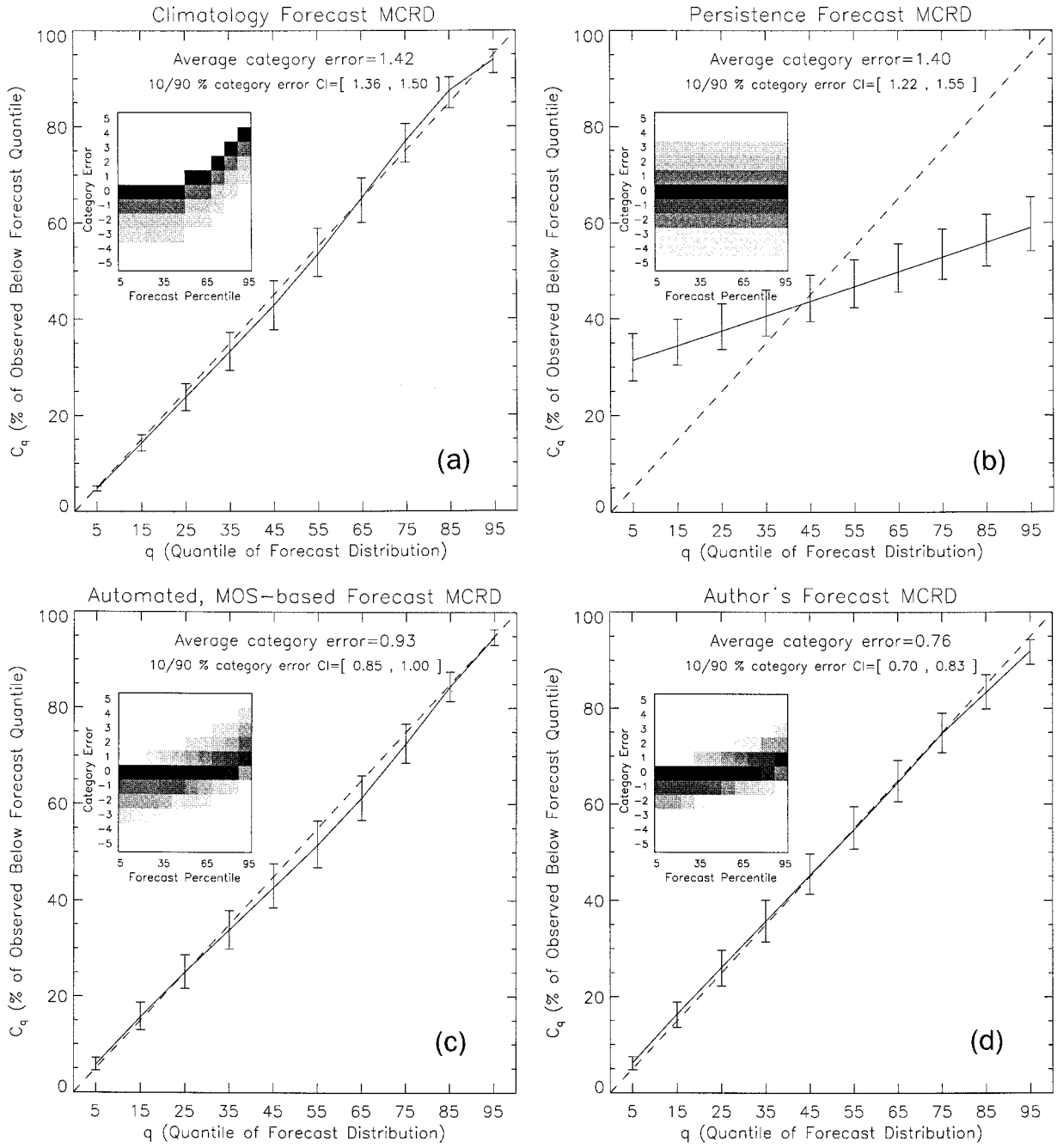

FIG. 3. Multicategory reliability diagram for five forecast contestants (a) climatology; (b) persistence; (c) automated, MOS-based scheme; (d) the author's subjective forecasts; and (e) forecast with dry bias. The inset box indicates the frequency of various category errors at each forecast percentile; the darker the box, the higher the percentage with that category error. Error bars indicate the 10th and 90th percentiles as determined through bootstrap testing.

than keeping track of the calibration at all quantiles, computations here will be performed only at preset quantiles at the middle of each 10% increment of the forecast. For example, the fifth quantile $q = .05$ will be used to represent the quantiles $.00 < q \leq .10$, and similarly $q = .15$ is used to represent $.10 < q \leq .20$; this is consistent with the initial simplification of rounding categorical probabilities to the nearest 10%. To reexpress the forecast vector in terms of category numbers, define a forecast category vector $\mathbf{z}_i$, $i = 1, \ldots, N$, composed here of 10 entries z_{iq}, where $q = \{0.05, 0.15, \ldots, 0.95\}$. A forecast vector $\mathbf{y}_i$ is converted into a vector $\mathbf{z}_i$ representing the forecast category number at each quantile. Hence, for the forecast of $\mathbf{y}_i = [0.7, 0.2,$

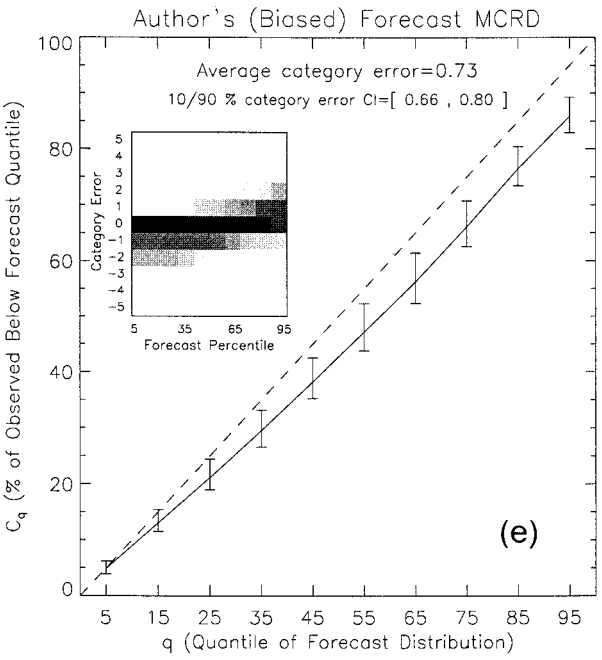

FIG. 3. *(Continued)*

0.1, 0, 0, 0], $z_{i.05} = \ldots = z_{i.65} = 1$, $z_{i.75} = z_{i.85} = 2$, and $z_{i.95} = 3$, or, $\mathbf{z}_i = [1, 1, 1, 1, 1, 1, 1, 2, 2, 3]$.

The calibration C_q for a given quantile q is the probability that the observed category o_i is less than the forecast category at this quantile, averaged over all N forecasts:

$$C_q = \overline{\mathrm{P}\{o_i < z_{iq}\}}. \tag{1}$$

For a single day's forecast at a given quantile, there are three possibilities: $o_i < z_{iq}$, $o_i = z_{iq}$, or $o_i > z_{iq}$. If $o_i < z_{iq}$, $\mathrm{P}(o_i < z_{iq}) = 1$, and similarly, if $o_i > z_{iq}$, $\mathrm{P}(o_i < z_{iq}) = 0$. If $o_i = z_{iq}$, then the probability is determined as follows. Suppose in the example above, the observed category was 2. Then, the 70th–90th percentile have the same category as the observation. For $0.0 < q \leq 0.70$, the forecast category is 1 and $o_i > z_{iq}$, so $\mathrm{P}(o_i < z_{iq}) = 0$. Similarly, $0.90 < q \leq 1.0$ the forecast category is 3 and $\mathrm{P}(o_i < z_{iq}) = 1$. Hence the probabilities 0 and 1 are boundary conditions at the 70th and 90th percentiles. Probabilities for quantiles in between vary linearly so perfect forecasts (all probability and observation in the same category) are perfectly calibrated. Let $q_{\min}$ and $q_{\max}$ be the lowest and highest quantiles where the observation and forecast have the same category. Then

$$P(o_i < z_{iq}) \begin{cases} = 0 \\ \quad \text{if } o_i > z_{iq} \\ = (q - q_{\min})/(q_{\max} - q_{\min}) \\ \quad \text{if } o_i = z_{iq} \\ = 1 \\ \quad \text{if } o_i < z_{iq}. \end{cases} \tag{2}$$

Figure 2 illustrates this.

The MCRD is generated by plotting C_q against q. Error bars are also plotted that represent the 10th and 90th percentiles of resampled multicategory reliabilities generated via a bootstrap test (Wilks 1995). The bootstrap used here was run 200 times.

Accurate calibration at each quantile is not a full indication of forecast quality. For example, forecasts may be nonspecific yet well calibrated; it is of course preferable to have a forecast that is both sharp and calibrated. Hence, included on the MCRD is a colored checkerboard plot of the forecast minus observed categories at the various quantiles. Darker colors on the checkerboard indicate more highly populated bins. A perfect forecast will have a black stripe at 0 category error and perfect calibration. Note that the checkerboard plot only indicates the category *differences* between forecasts and observations; the observation categories are never indicated. Thus, the MCRD does not illustrate the full complexity of the joint forecast/observation distributions.

As a summary of the checkerboard plot, the mean absolute category error between forecast and observed is noted, averaged over all forecasts and all quantiles. Further, the 10th and 90th percentiles of the average category error are also given to indicate the range of uncertainty. These were also generated from a 200-sample bootstrap test.

3. Demonstration using forecast contest data

Using precipitation forecasts generated in the Cornell University forecast contest during the academic year 1995/96, the MCRD is demonstrated. The next day's total 24-h rainfall was predicted on a total of $N = 124$ days. Data from five ''contestants'' will be shown: climatology; persistence; an automated, MOS-based forecast; the author's subjective forecasts; and a forecast with a dry bias. For all forecasts, probabilities were set for the MOS categories described in section 1, and probabilities for each category were rounded to the nearest 10th. The climatology forecast is a generated from the distribution of observed categories over the 124 days. The actual distribution was $\mathbf{y}_i \approx [0.48, 0.23, 0.12, 0.09, 0.04, 0.04]$. Since forecasts must be rounded to the nearest 10th, this was changed to $\mathbf{y}_i = [0.5, 0.2, 0.1, 0.1, 0.1, 0]$. The MOS-based forecasts combined the 12-h probability of precipitation (PoP) forecasts into a single 24-h PoP (Wilks 1990a), and then assessed probabilities for nonzero precipitation amount categories using the PoP and gamma distributions fit to conditional climatologies (Wilks 1990b). The biased forecasts were created by adjusting the author's subjective forecasts so that the forecast categories above the fifth percentile were shifted by 10%. For example, in each forecast, $z_{i.15}$ is shifted up to $z_{i.25}$, $z_{i.25}$ to $z_{i.35}$, and so on.

Reliability diagrams for each contestant are shown in Figs. 3a–e. The climatology forecast in Fig. 3a appears well calibrated; deviations from calibration are not sig-

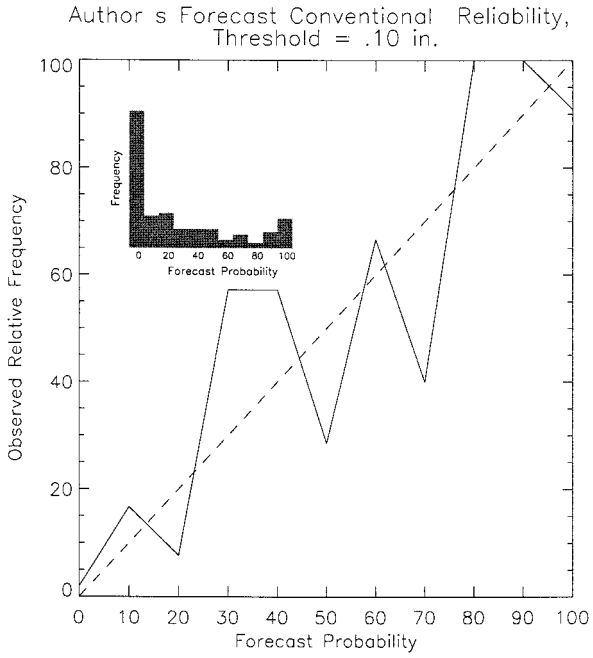

FIG. 4. Conventional reliability diagram for the author's precipitation forecasts, with a threshold of 0.10 in. (2.54 mm).

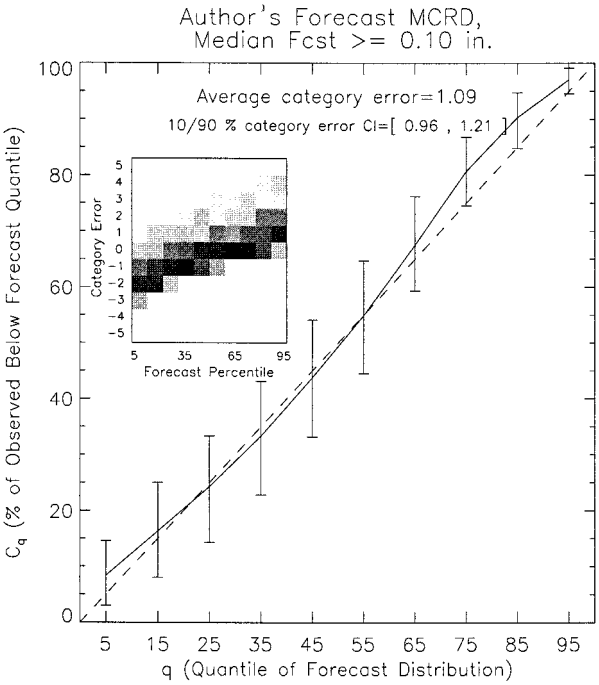

FIG. 5. Multicategory reliability diagram for the subset of the author's precipitation forecasts where the median forecast was greater than or equal to 0.10 in.

nificant and are likely due to the truncation of the climatology distribution to the nearest 10th.

Consider next the persistence forecast in Fig. 3b. Here, 100% of the probability is allocated to the category in which the previous day's precipitation amount occurred. Hence, persistence forecast categories are identical across all quantiles. As shown, its reliability curve is too flat; the forecast categories for the lower quantiles were set too high on average. The opposite is true at the higher quantiles, where far fewer than expected observation categories were lower than the forecast category. Also, as noted in the checkerboard plot, there were many forecasts with multicategory errors, as indicated by the relatively dark shades assigned to nonzero category errors and the average category error of 1.40. Since each day's category forecast was the same for all quantiles, the checkerboard plot of category errors appears as horizontal bands.

The automated, MOS-based forecasts (Fig. 3c) and the author's forecasts (Fig. 3d) are both well calibrated. However, as shown by the lower average category error and the more confined shades on the checkerboard plot, the author's subjective forecasts tend to be sharper and lower in error than those from the automated forecasts. Both have lower category errors than climatology or persistence.

The biased forecasts are not well calibrated, as expected. Fewer than expected observations were in categories less than the forecast category, especially at higher quantiles. To remedy this, more probability needs to be allocated to the higher categories, so that the quantiles of the probability distribution are shifted up to higher categories. This makes sense, as the artificial biased forecasts were created through a reverse process, described earlier. In general, when the reliability line is above the diagonal, there is a moist bias at these quantiles; when below the diagonal, a dry bias.

The usefulness of MCRDs is illustrated through a comparison with a conventional reliability diagram. For example, the author wished information on the errors in the forecast probability distribution for events with significant precipitation, category 3 and greater (0.10 in. and above). A conventional reliability diagram for the author's forecasts at the 0.10-in. threshold is presented in Fig. 4. As shown, the information that can be gleaned from the reliability diagram is nil because of the inadequate sample size for many of the forecast probabilities ($N = 124$, and the majority of the forecasts at 0% probability of greater than 0.10 in.). Compare this with Fig. 5, an MCRD for the subset of forecasts where the median forecast ($q_{.5}$) was greater than 0.10 in. Here there are $N = 27$ forecasts at *each quantile*, making inferences much more trustworthy because of the greater sample size. As shown, at the highest quantiles there is a moist bias, though the deviation from perfect calibration is barely significant at the 10% level. This suggests that if future categorical forecast distributions have similar characteristics, they can be adjusted in subsequent forecasts by slightly shifting the forecast probability distribution. By allocating less probability to higher precipitation categories, the right tail of the distribution is lessened and the forecast categories for higher quantiles

are lowered. This will have the effect of lessening the number of observations that are lower than the threshold at this quantile, achieving better calibration.

4. Conclusions

The extension of the reliability diagram to multiple-category probabilistic forecasts is presented here. This new diagram is shown to be useful for evaluating some errors in the assessment of probability distributions in such forecasts, though it cannot replace the RPSS for ranking contestants nor does it provide a complete description of the forecast error. Despite this, the MCRD works well even for relatively small sample sizes, when conventional reliability diagrams are inadequate.

Acknowledgments. This research was supported under NSF Grant ATM-9508645. The author thanks Caren Marzban and Harald Daan for constructive criticism that markedly improved this manuscript.

REFERENCES

Daan, H., 1985: Sensitivity of the verification scores to classification of the predictand. *Mon. Wea. Rev.*, **113,** 1384–1392.

Dallavalle, J. P., J. S. Jensenius Jr., and S. A. Gilbert, 1992: NGM-based MOS guidance—The FOUS14/FWC message. Technical Procedures Bull. 408, NOAA/National Weather Service, Washington, DC, 9 pp. [Available from NOAA/NWS, Services Development Branch, 1325 East–West Highway, Room 13466, Silver Spring, MD 20910.]

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8,** 985–987.

Hamill, T. M., and D. S. Wilks, 1995: A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Wea. Forecasting,* **10,** 620–631.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10,** 155–156.

Wilks, D. S., 1990a: On the combination of forecast probabilities for consecutive precipitation periods. *Wea. Forecasting,* **5,** 640–650.

——, 1990b: Probabilistic quantitative precipitation forecasts derived from PoPs and conditional precipitation amount climatologies. *Mon. Wea. Rev.*, **118,** 874–882.

——, 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.